Analysis of a Discrete-Time Queueing System with an *NT*-Policy

Bart Feyaerts, Stijn De Vuyst, Sabine Wittevrongel, Herwig Bruneel

SMACS Research Group, TELIN Department, Ghent University
Sint-Pietersnieuwstraat 41, 9000 Gent, Belgium
E-mail: {bfeyaert, sdv, sw, hb}@telin.ugent.be

Abstract. In this paper, we analyse a discrete-time single-server queue operating under the NT-policy, which aims at clustering the service of customers in order to reduce the number of server activations and deactivations. Starting from an empty queue, the service of arriving customers is postponed until either of two thresholds is reached. Specifically, exhaustive service of customers is initiated only if either N customers have accumulated (space threshold) *or* if more than T slots have passed since the arrival of the first customer. This way, the queue cycles between three states, i.e. an empty phase, an accumulating phase and a serving phase. A Bernoulli arrival process and deterministic service times are assumed. We derive the steady-state probabilities of the system's state as well as the distributions of the phase sojourn times and the customer delay. For the latter, we condition on the phase during the customer's arrival slot. The influence of the model parameters on the results is discussed by means of a numerical example.

1 Introduction

In a typical work-conserving queue under low to moderate load conditions, the service unit has to switch often between being idle and being busy. This frequent activation and/or deactivation of the service unit may pose a considerable overhead, e.g. with machines that need to power up, be configured, checked or undergo any other costly initialisation procedure before customers can be served after a period of idleness. In such cases, it is beneficial to *cluster* the customer services to some extent by using a threshold policy such as the N-policy, first presented in [1]. Under this policy the server is deactivated if the queue is depleted as usual but is only reactivated once $N > 1$ customers have accumulated again, instead of only one in the work-conserving case. This assures longer uninterrupted busy periods, such that less server switch-overs are required. Since [1], many adaptations of this N-policy have been proposed and studied in literature. Up to recently, most of this research [2–4] has been done in a continuous-time setting while far less attention has been spent on discrete-time models. Nevertheless, batch arrivals and batch service for discrete-time N-policy queues are studied in [5]. In [6], a bilevel threshold mechanism is studied and in [7], service is differentiated between the N accumulated customers and later arrivals.

Although the N-policy effectively reduces server switch-overs it also increases the queueing time (delay) of the customers. Consider in particular a first customer arriving when the queue is empty, then its service is delayed until such time as $N - 1$ *other* customers have arrived as well. Clearly, if the arrival rate is very low this may result in customer starvation, i.e. the customer delay tends to infinity. The NT-policy counters this drawback by imposing a time limit T on the accumulation time, besides the space threshold N. So, the server reactivates when the queue has length N *or* if the first customer has been waiting in the queue for a time T, whichever happens first. Continuous-time models of this double threshold policy are found in [8–10]. In this paper however, we propose an analysis in a discrete-time setting.

The paper is organised as follows. In Sect. 2, we present a mathematical model of the NT-policy. This model is then used in Sect. 3 to analyze the system's behaviour. The analysis allows us to determine some interesting and useful measures in Sects. 4–5. Sect. 6 is focused on the delay performance of the NT-policy. We then illustrate the properties of the NT-policy in Sect. 7 with some numerical results and compare to the N-policy in Sect. 8. Finally, Sect. 9 concludes this paper.

2 Model Description

We consider a discrete-time single-server queue with infinite storage capacity operating under the NT-policy. Time is divided into fixed-length intervals called *slots*, corresponding to the service time required by a single costumer. The arrivals of customers form a Bernoulli process with rate λ, such that in each slot a customer arrives with probability λ and no customer arrives with probability $1-\lambda$. The number of arrivals during slot k is referred to as a_k. Thus, the system load ρ equals the arrival rate λ and stability is assured, even if $\lambda = 1$.

The NT-policy implies that when the server becomes idle, it deactivates and will remain inactive until exactly N customers have accumulated in the queue and/or until there is a customer in the queue for exactly T slots. Note that only situations where $1 < N \leq T$ are of interest to us. Indeed, for $N = 1$ the policy is the same as in a normal work-conserving queueing system. If on the other hand $T < N$, only the time threshold T would be relevant since it takes at least N slots for N customers to accumulate in the queue. This system would therefore only implement a T-policy. Hence we will restrict the analysis to systems where the inequality $1 < N \leq T$ holds. Note that if T tends to infinity, the system converges to an N-policy system; if N tends to infinity as well, the system will never be reactivated once it has become empty.

Due to the NT-policy, the system's operation exhibits a cyclic behaviour, as illustrated in Fig. 1. When a first customer arrives in an empty system, the system proceeds to an accumulating state until at least one of the thresholds is reached. Thereupon the system will start serving the customers exhaustively until it becomes empty again. Thus, we distinguish three subsequent phases, i.e.

empty, *accumulating* customers and *serving* customers. The total time for the system to complete all three phases, is referred to as a cycle with length Q.

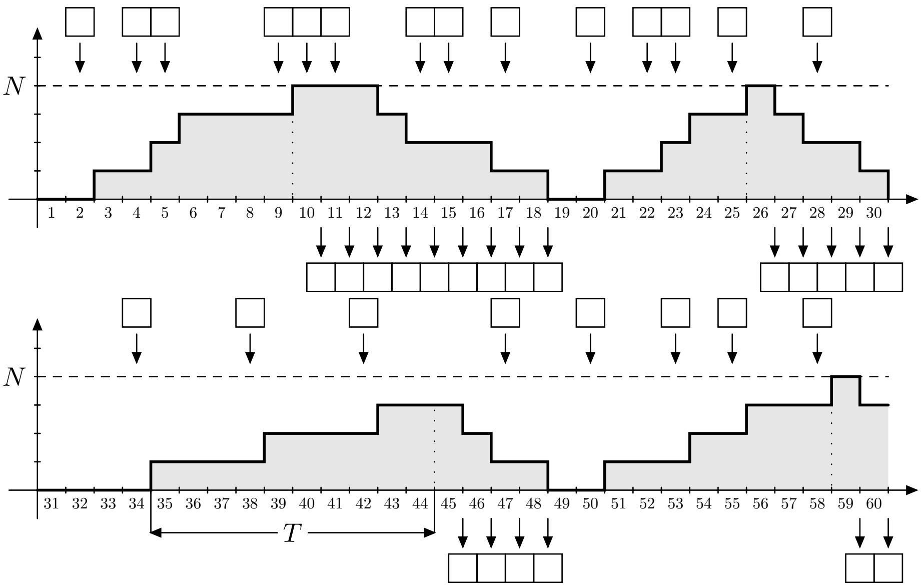

Fig. 1. Evolution of the system content of an NT-policy system with $N = 4$ and $T = 10$. The vertical dotted lines denote the transition from an accumulating phase to a serving phase.

3 System Equations and Buffer Analysis

In order to analyse the behaviour of the NT-policy system, we first introduce the random variable ϕ_k as the phase in which the system resides during slot k. This variable can take only the values 0, 1 and 2 to refer to the empty phase, the accumulating phase and the serving phase respectively. In what follows, we will mainly refer to the different phases by using their corresponding index.

We also introduce the random variable t_k to represent the sojourn time of the first customer in the queue at the end of a random slot k in phase 1. Specifically, if a first customer arrives in an empty queue during slot $k-1$, the phase in slot k becomes 1 and the variable t_k takes value 1. It is clear that $1 \leq t_k \leq T$ for any slot k in phase 1. If $t_k = T$, then for sure the system proceeds to phase 2 in slot $k+1$. For simplicity, we assume $t_k = 0$ for any slot k in phase 0 or in phase 2.

Finally, we introduce the random variable u_k as the system content at the beginning of slot k, this is before any arrivals or departures.

How the system evolves from slot to slot, is described by the system equations (1)–(3), depending on the value of ϕ_k.

$\phi_k = 0$	$\phi_k = 1$	$\phi_k = 2$	
$u_{k+1} = a_k$	$u_k + a_k$	$u_k + a_k - 1$	(1)
$t_{k+1} = \begin{cases} 0, & a_k = 0 \\ 1, & a_k > 0 \end{cases}$	$\begin{cases} t_k + 1, & t_k < T \wedge u_{k+1} < N \\ 0, & t_k = T \vee u_{k+1} = N \end{cases}$	0	(2)
$\phi_{k+1} = \begin{cases} 0, & a_k = 0 \\ 1, & a_k > 0 \end{cases}$	$\begin{cases} 1, & t_k < T \wedge u_{k+1} < N \\ 2, & t_k = T \vee u_{k+1} = N \end{cases}$	$\begin{cases} 2, & u_{k+1} > 0 \\ 0, & u_{k+1} = 0 \end{cases}$	(3)

The system equations show that the set of vectors $\{(\phi_k, t_k, u_k)\}$ forms a Markov chain. Therefore the vector (ϕ_k, t_k, u_k) is sufficient to describe the system state at a random slot k and as such, it is called the system state vector.

The next step in the analysis is to introduce the following probabilities:

$$
\begin{aligned}
p_0 &\triangleq \text{Prob}[\phi_k = 0] \ , && && (4)\\
p_{1,m,n} &\triangleq \text{Prob}[\phi_k = 1, t_k = m, u_k = n] \ , && 1 \le n \le N-1, n \le m \le T \ , && (5)\\
p_{2,n} &\triangleq \text{Prob}[\phi_k = 2, u_k = n] \ , && 1 \le n \le N \ . && (6)
\end{aligned}
$$

We will not determine expressions for p_0, $p_{1,m,n}$ and $p_{2,n}$ directly, rather we will determine the corresponding coefficients q_0, $q_{1,m,n}$ and $q_{2,n}$ defined as

$$
q_0 \triangleq \frac{p_0}{p_{1,1,1}} \ , \qquad q_{1,m,n} \triangleq \frac{p_{1,m,n}}{p_{1,1,1}} \ , \qquad q_{2,n} \triangleq \frac{p_{2,n}}{p_{1,1,1}} \ . \tag{7}
$$

The idea is that $p_{1,1,1}$ corresponds to an event that occurs precisely once per cycle; i.e. it refers to the first slot of the accumulating phase. Therefore the q's correspond to the fraction of certain events within a single cycle.

First, we find that

$$
p_{1,1,1} = \lambda p_0 \Leftrightarrow q_0 = \frac{1}{\lambda} \ . \tag{8}
$$

This can be understood by the fact that the system shifts from the empty phase to the accumulating phase under influence of an arriving customer. For $q_{1,m,n}$, we first notice that in phase 1, the value of t_k increments with 1 and u_k increments with the number of arrivals during slot k. Thus we find

$$
\begin{aligned}
q_{1,m,n} &= (1-\lambda) q_{1,m-1,n} + \lambda q_{1,m-1,n-1} \\
&= \binom{m-1}{n-1} \lambda^{n-1} (1-\lambda)^{m-n} \ , \qquad 1 < n \le m \ ,
\end{aligned} \tag{9}
$$

with $q_{1,m,1} = (1-\lambda)^{m-1}, \forall m \ge 1$. Indeed, after the arrival of the first customer in an empty system, it takes $m-1$ slots with a total of $n-1$ arrivals to end up in phase 1 with a sojourn time m and buffer content n. Note that the value of any $q_{1,m,n}$ is independent of both thresholds N and T, it only depends on m, n and λ. The coefficient $q_{2,N}$ corresponds to a system with N customers in phase

2. This can only occur if there was an arrival and either the system was already in phase 2 or the system was in phase 1 and the N-threshold has been reached:

$$\begin{aligned} q_{2,N} &= \lambda \sum_{m=N-1}^{T} q_{1,m,N-1} + \lambda q_{2,N} = \frac{\lambda}{1-\lambda} \sum_{m=N-1}^{T} q_{1,m,N-1} \\ &= \sum_{m=N-1}^{T} \binom{m-1}{N-2} \lambda^{N-1} (1-\lambda)^{m-N} \quad . \end{aligned} \tag{10}$$

An expression for $q_{2,n}, n < N$ can then be found by expressing that the corresponding event could have been reached either from within phase 2 or from phase 1 in case the T-threshold has been reached. So we find

$$\begin{aligned} q_{2,n} &= (1-\lambda) q_{1,T,n} + \lambda q_{1,T,n-1} + (1-\lambda) q_{2,n+1} + \lambda q_{2,n} \\ &= \frac{1}{1-\lambda} q_{1,T+1,n} + q_{2,n+1} = \ldots = \frac{1}{1-\lambda} \sum_{j=n}^{N-1} q_{1,T+1,j} + q_{2,N} \\ &= \sum_{j=n}^{N-1} \binom{T}{j-1} \lambda^{j-1} (1-\lambda)^{T-j} + \sum_{m=N-1}^{T} \binom{m-1}{N-2} \lambda^{N-1} (1-\lambda)^{m-N} \quad . \end{aligned} \tag{11}$$

Note that in the above we have used $q_{1,T+1,n} \triangleq (1-\lambda) q_{1,T,n} + \lambda q_{1,T,n-1}$. Although this corresponds to an event that is impossible to occur in the current system, the expression itself is justified.

From (7), the probabilities p_0, $p_{1,m,n}$ and $p_{2,n}$ can now be written in terms of $p_{1,1,1}$. The latter probability can then finally be found from the normalization condition:

$$1 = p_0 + \sum_{n=1}^{N-1} \sum_{m=n}^{T} p_{1,m,n} + \sum_{n=1}^{N} p_{2,n} \quad ,$$

which leads to

$$p_{1,1,1} = \left(q_0 + \sum_{n=1}^{N-1} \sum_{m=n}^{T} q_{1,m,n} + \sum_{n=1}^{N} q_{2,n} \right)^{-1} \quad . \tag{12}$$

In the Sects. 4 and 5, the obtained distribution of the Markovian system state will be used to study the distribution of the phase and cycle durations and the probabilities of being in a certain phase. These in turn will then enable us to study the delay of a customer under the NT-policy in Sect. 6.

4 Phase and Cycle Duration

As pointed out earlier, the system exhibits a cyclic behaviour, with each cycle consisting of the three subsequent phases 0, 1 and 2. In this section, we will derive expressions for the phase sojourn times, this is the number of slots the system resides in a certain phase. We introduce Φ_i as the phase i sojourn time, with probability generating function (pgf) $\Phi_i(z)$ ($i \in \{0, 1, 2\}$).

Empty phase. As can be derived from the system equations, the empty phase kicks in as soon as the system becomes empty, and ends at the end of the first slot with an arrival. An empty phase of t slots $(1 \leq t)$ can therefore only occur if there are $t-1$ consecutive slots without any arriving customers, followed by a slot during which a customer does arrive. From this notion, we see that

$$\text{Prob}[\Phi_0 = t] = \lambda(1-\lambda)^{t-1} \ \ , 1 \leq t \ \ , \tag{13}$$

$$\Phi_0(z) \triangleq E\big[z^{\Phi_0}\big] = \frac{\lambda z}{1-(1-\lambda)z} \ \ . \tag{14}$$

Accumulating phase. At the beginning of the first slot of any accumulating phase, there is exactly one customer in the queue, and its sojourn time at the end of the slot is exactly 1. An accumulating phase ends either when the Nth customer arrives at the queue or when the first customer has been waiting for T slots, or both, whichever occurs first. The accumulating phase sojourn time will then be T, unless the Nth customer arrives at the queue sooner. This can be expressed as

$$\text{Prob}[\Phi_1 = t] = \begin{cases} \lambda q_{1,t,N-1} \ \ , & N-1 \leq t \leq T-1 \ \ , \\ \sum_{n=1}^{N-1} q_{1,T,n} \ \ , & t = T \ \ , \end{cases} \tag{15}$$

with pgf

$$\Phi_1(z) \triangleq E\big[z^{\Phi_1}\big] = \lambda \sum_{m=N-1}^{T-1} q_{1,m,N-1} z^m + \sum_{n=1}^{N-1} q_{1,T,n} z^T \ \ . \tag{16}$$

We now introduce ω as

$$\omega \triangleq \text{Prob}[N \text{ customers have accumulated during } \Phi_1] = \lambda \sum_{t=N-1}^{T} q_{1,t,N-1} \ \ . \tag{17}$$

Serving phase. During the final phase the server is active and the customers get served. The phase lasts until the queue becomes empty and no more customers are present in the system. The phase 2 duration is the result of two aspects: the number of customers in the queue at the start of the phase and the number of new customer arrivals during the phase. Therefore we first introduce Δ as the time that is needed to reduce the number of customers in the system by 1. The total number of customers in the system only decreases when no new customers arrive, whereas it remains unaltered when there is an arrival. Therefore the distribution of Δ can be found as $\text{Prob}[\Delta = t] = \lambda^{t-1}(1-\lambda), 1 \leq t$, with pgf

$$\Delta(z) \triangleq E\big[z^{\Delta}\big] = \frac{(1-\lambda)z}{1-\lambda z} \ \ . \tag{18}$$

The number of customers in the system at a phase 2 start is highly dependent on the preceding accumulating phase. If the corresponding $\Phi_1 < T$ we know for sure that there are N customers in the queue, whereas for $\varPhi_1 = T$ there could be less customers. This is reflected in the joint pgf $\varPhi_{1,2}(x,y)$ of $\varPhi_1$ and $\varPhi_2$ as

$$\begin{aligned}\varPhi_{1,2}(x,y) \triangleq E\big[x^{\varPhi_1} y^{\varPhi_2}\big] &= \lambda \varDelta(y)^N \sum_{m=N-1}^{T-1} q_{1,m,N-1} x^m \\ &\quad + \sum_{n=1}^{N-1} q_{1,T,n} x^T \left((1-\lambda)\varDelta(y)^n + \lambda \varDelta(y)^{n+1}\right) \ . \end{aligned} \tag{19}$$

The pgf $\varPhi_2(z)$ of $\varPhi_2$ can then be found by substituting $x = 1$ and $y = z$ in (19); this yields

$$\varPhi_2(z) = \lambda \varDelta(z)^N \sum_{m=N-1}^{T-1} q_{1,m,N-1} + \sum_{n=1}^{N-1} q_{1,T,n} \left((1-\lambda)\varDelta(z)^n + \lambda \varDelta(z)^{n+1}\right) \ . \tag{20}$$

Cycle length. The total length Q of an arbitrary cycle can then be found as the sum of the lengths of the three constituting phases. Note that, especially for $\varPhi_1$ and $\varPhi_2$, we must consider consecutive phases, such that the number of customers accumulated in phase 1 corresponds to the initial number of custumers of phase 2. The pgf $Q(z)$ of the cycle length is then given by

$$\begin{aligned} Q(z) &\triangleq E\big[z^{\varPhi_0+\varPhi_1+\varPhi_2}\big] = \varPhi_0(z)\varPhi_{1,2}(z,z) \\ &= \frac{\lambda z}{1-(1-\lambda)z} \Bigg(\lambda \varDelta(z)^N \sum_{m=N-1}^{T-1} q_{1,m,N-1} z^m \\ &\qquad + \sum_{n=1}^{N-1} q_{1,T,n} z^T \left((1-\lambda)\varDelta(z)^n + \lambda \varDelta(z)^{n+1}\right) \Bigg) \ . \end{aligned} \tag{21}$$

5 Phase Probability

Analogously to p_0 earlier, we introduce p_1 and p_2, such that $p_i \triangleq \mathrm{Prob}[\phi_k = i]$, $i \in \{0,1,2\}$. The p_i can be understood as the fraction of the time the system is in phase i, and therefore it can be seen that $p_i \triangleq \frac{E[\varPhi_i]}{E[Q]}$, $i \in \{0,1,2\}$. Specifically for p_0 we find that

$$p_0 = \frac{E[\varPhi_0]}{E[Q]} = \frac{1}{\lambda E[Q]} \ , \tag{22}$$

or from (8)

$$E[Q] = \frac{1}{\lambda p_0} = \frac{1}{p_{1,1,1}} = q_0 + \sum_{n=1}^{N-1} \sum_{m=n}^{T} q_{1,m,n} + \sum_{n=1}^{N} q_{2,n} \ . \tag{23}$$

Also p_2 can be determined in an atypical way. As mentioned before, we assume the system load to be less then one so that the system is stable. In equilibrium the mean arrival rate must be equal to the mean departure rate. Since departures only occur in phase 2 and that at a rate of 1 departure per slot, the mean departure rate is equal to p_2. So we find

$$p_2 = \lambda \ . \tag{24}$$

From (22) and (24) and the normalization condition, we get

$$p_1 = 1 - p_0 - p_2 \ . \tag{25}$$

6 Customer Delay Distribution

The customer delay is the integer number of slots a customer resides in the system, starting at the end of the customer's arrival slot, until the end of the slot during which the customer leaves the system. In this section we will pick a random customer $\mathcal{C}$, and determine what delay this customer suffers. As can be expected, the customer delay will highly depend on the system state at the beginning of the customer's arrival slot; for convenience we will refer to this arrival slot as slot I. The BASTA property (Bernoulli Arrivals See Time Averages) [11] assures that the distribution of the system state in $\mathcal{C}$'s arrival slot I is the same as the system state distribution during a random slot, even though I is not a random slot itself. This notion is essential to the delay analysis presented here.

In what follows we will determine the delay d_i for a random customer $\mathcal{C}$ that arrives during phase i of a cycle ($i \in \{0, 1, 2\}$). Finally we will combine these results to find the delay d of a random customer $\mathcal{C}$, regardless of the phase during which $\mathcal{C}$ enters the system.

$\mathcal{C}$ arrives during phase 0. Any customer that arrives during an empty phase, will trigger an accumulating phase at the beginning of the next slot. When the accumulating phase ends, this customer will be the first to get served. Thus the delay of customer $\mathcal{C}$ that arrives during phase 0 consists of the entire induced phase 1 and $\mathcal{C}$'s service time, or $d_0 = \Phi_1 + 1$ with pgf

$$D_0(z) \triangleq E\left[z^{d_0}\right] = z\Phi_1(z) \ . \tag{26}$$

Fig. 2 shows how $\mathcal{C}$ (the dark square) passes through the buffer.

$\mathcal{C}$ arrives during phase 1. If a customer $\mathcal{C}$ arrives during an accumulating phase, $\mathcal{C}$ will have to wait until the phase ends and all previously arrived customers have left the system before the server is ready to take care of $\mathcal{C}$ as is depicted in Fig. 3. Note that at the beginning of slot I, exactly u_I customers are present in the system and by the end of slot I the first customer has been waiting for exactly t_I slots. The current accumulation phase then lasts until, starting from slot $I+1$, $N - u_I - 1$ more customers have accumulated or $T - t_I$ more slots have passed, whichever happens first. This time span corresponds to an accumulating phase duration of an $N'T'$-policy system with $N' = N - u_I$ and $T' = T - t_I$.

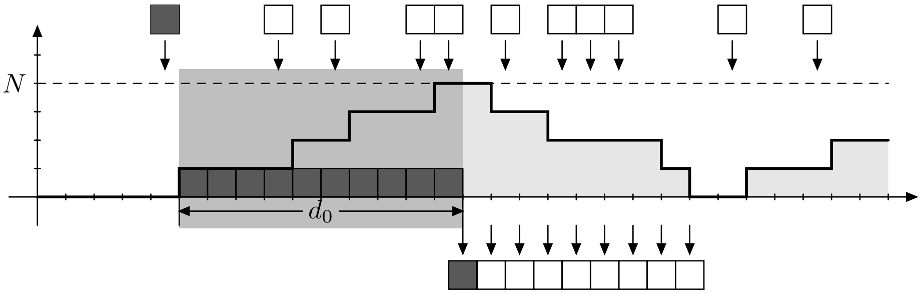

Fig. 2. Graphic representation of d_0.

Therefore the delay of $\mathcal{C}$ can be expressed as $d_1 = \Phi_1^{(N',T')} + u_I + 1$, where we introduced $\Phi_1^{(N',T')}$ as the phase 1 sojourn time of the $N'T'$-policy system. In case $1 < N' \leq T'$, we may use (15)–(16) directly to obtain the distribution of $\Phi_1^{(N',T')}$, only by substituting the alternate values for the thresholds. Indeed the values for the different $q_{1,m,n}$ remain unchanged, since they are independent of both the thresholds. Note however that if $u_I = N-1$ or $t_I = T$, (15)–(16) no longer hold; in these cases the accumulating phase ends immediately after slot I, and $\Phi_1^{(N',T')}$ in the expression for d_1 should be taken equal to 0. The distribution of d_1 is then given by

$$\begin{aligned}
&\mathrm{Prob}[d_1 = t] = \mathrm{Prob}\Big[\Phi_1^{(N',T')} = t - u_I - 1\Big] \\
&= \sum_{n=1}^{N-1} \sum_{m=n}^{T} \frac{p_{1,m,n}}{p_1} \mathrm{Prob}\Big[\Phi_1^{(N-n,T-m)} = t - n - 1\Big] \\
&= \frac{1}{p_1} \left(\sum_{n=1}^{(N-2,t-2)^-} p_{1,T-t+n+1,n} \sum_{i=1}^{N-n-1} q_{1,t-n-1,i} + \left[\sum_{m=N-1}^{T-1} p_{1,m,N-1} \right]_{t=N} \right. \\
&\quad \left. + \left[p_{1,T,t-1}\right]_{2\leq t\leq N} + \left[\lambda \sum_{n=1}^{N-2} \sum_{m=n}^{T-t+n} p_{1,m,n} q_{1,t-n-1,N-n-1} \right]_{t\geq N} \right) , \qquad (27)
\end{aligned}$$

where we introduced $(x,y)^-$ as the minimum of x and y and also $[x]_c$ to be equal to x if the condition c holds, and to be equal to 0 otherwise. The pgf $D_1(z)$ of d_1 is then given by

$$\begin{aligned}
&D_1(z) \triangleq E\big[z^{d_1}\big] \\
&= \frac{1}{p_1} \left(\sum_{t=2}^{T+1} z^t \sum_{n=1}^{(N-2,t-2)^-} p_{1,T-t+n+1,n} \sum_{i=1}^{N-n-1} q_{1,t-n-1,i} + z^N \sum_{m=N-1}^{T-1} p_{1,m,N-1} \right. \\
&\quad \left. + \sum_{t=2}^{N} p_{1,T,t-1} z^t + \lambda \sum_{t=N}^{T+1} z^t \sum_{n=1}^{N-2} \sum_{m=n}^{T-t+n} p_{1,m,n} q_{1,t-n-1,N-n-1} \right) . \qquad (28)
\end{aligned}$$

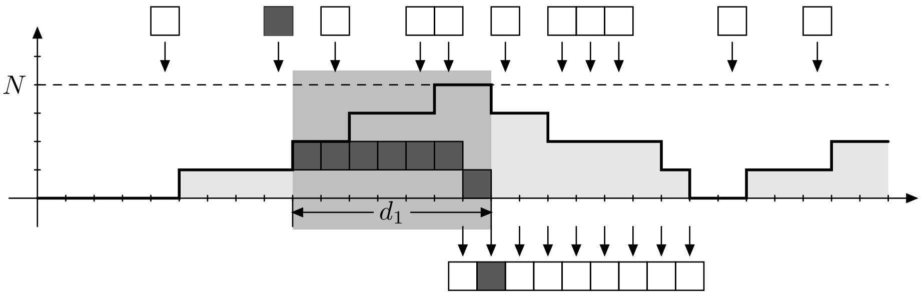

Fig. 3. Graphic representation of d_1.

$\mathcal{C}$ arrives during phase 2. During phase 2 the server is busy serving customers. Therefore the only delay suffered by a customer $\mathcal{C}$ arriving in a phase 2 slot, is the time needed to serve all customers (including $\mathcal{C}$) in the queue at the end of slot I. Since one customer gets served during this slot, we have $d_2 = u_I$ with pgf

$$D_2(z) \triangleq E\big[z^{d_2}\big] = \sum_{n=1}^{N} \frac{p_{2,n}}{p_2} z^n = \frac{p_{1,1,1}}{\lambda} \sum_{n=1}^{N} q_{2,n} z^n \quad . \tag{29}$$

In Fig. 4, we see how $\mathcal{C}$ is inserted in the queue, comes closer to the server and eventually leaves the system.

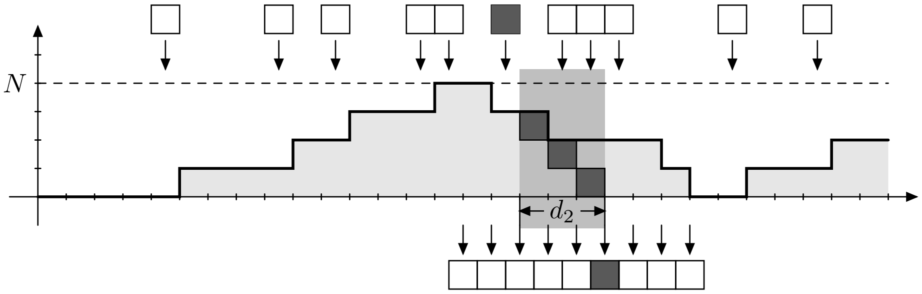

Fig. 4. Graphic representation of d_2.

Customer delay. The delay distribution of a random customer $\mathcal{C}$, regardless of the phase during which $\mathcal{C}$ enters the system, can then be found as

$$\mathrm{Prob}[d = t] = p_0\mathrm{Prob}[d_0 = t] + p_1\mathrm{Prob}[d_1 = t] + p_2\mathrm{Prob}[d_2 = t] \quad , \tag{30}$$

with pgf

$$D(z) \triangleq E\left[z^d\right] = p_0 D_0(z) + p_1 D_1(z) + p_2 D_2(z) \ . \tag{31}$$

7 Numerical Results

In this section, we will concentrate on some numerical results of the system's characteristics. The examples here are based on an NT-policy system, with $N = 40$ and $T = 100$. Unless otherwise stated, we consider an arrival rate $\lambda = 0.4$. In this configuration $N = \lambda T$ and hence, the mean number of slots needed to accumulate N customers equals T.

First we consider the mean phase sojourn times versus the arrival rate λ, depicted in Fig. 5. As one could expect from (13), the mean empty phase sojourn time $(\frac{1}{\lambda})$ decreases when λ increases. Also, we notice that for $\lambda \leq 0.4$, the mean phase 1 sojourn time is virtually equal to 100. This is due to the fact that for low arrival rates, the threshold T will generally be the one that triggers a new phase 2. Since every arrival during a serving phase prolongs the phase's sojourn time, the mean phase 2 sojourn time increases for increasing arrival rates.

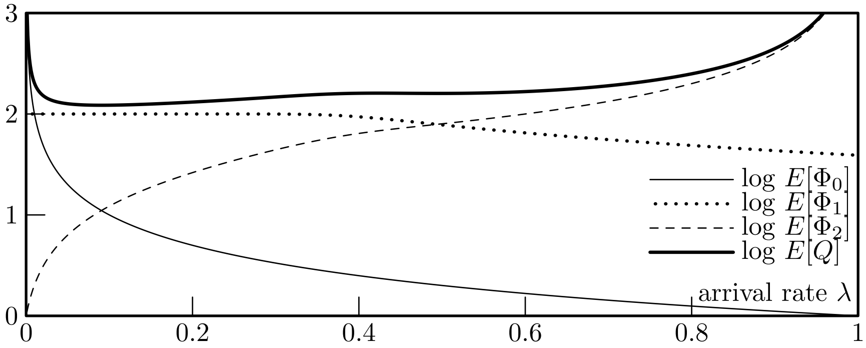

Fig. 5. Mean phase sojourn times $E[\Phi_i]$ $(i \in \{0,1,2\})$ and mean cycle length $E[Q]$ in slots versus the arrival rate λ on a log scale for $N = 40$ and $T = 100$.

Fig. 6 shows the effect of the arrival rate λ on the phase probabilities p_i $(i \in \{0,1,2\})$. These probabilities are important as they serve as a weight factor in the calculation of the customer delay distribution and the mean customer delay, as shown in (30). We see that the probability of a random slot to be part of an empty phase is high for extremely low arrival rates, but it very soon drops as the arrival rate gets higher; for the greater part of the graph, p_0 is even negligible. The probability p_1 runs a very different course: if the arrival rate is very low, the majority of time will be spent in phase 0, but as the arrival rate increases, the server will be empty less and both p_1 and p_2 will increase. Since the phase 1 sojourn times are limited to a maximum of T slots, only the serving phase will become longer due to an increasing arrival rate, therefore p_1 will decrease again, while p_2 continues to rise linearly.

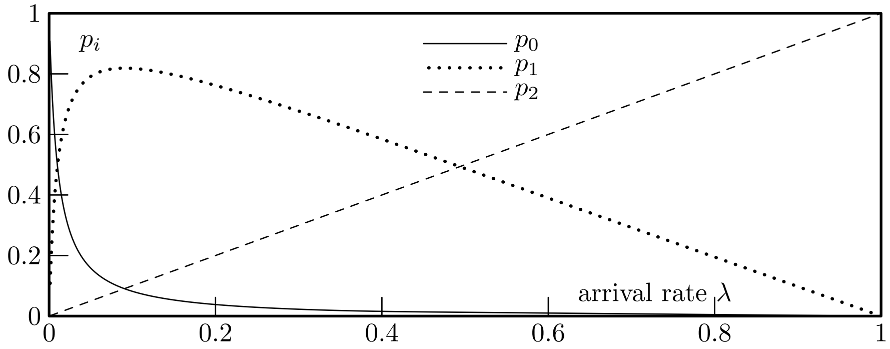

Fig. 6. Phase probabilities p_i $(i \in \{0, 1, 2\})$ versus the arrival rate λ for $N = 40$ and $T = 100$.

We now concentrate on the customer delay distribution, as presented in Fig. 7. Note the limited support of the different stochastic variables d_i $(i \in \{0, 1, 2\})$ and how the three corresponding curves are very different but add up to a suprisingly smooth curve for d, with only two outliers. The outlier at $t = T+1 = 101$ originates from the distribution of d_0 and corresponds to all cycles where the threshold T is reached at the end of phase 1. Thus, the sudden peak accumulates all cycles where the timer expires, regardless of how many customers eventually did arrive during the accumulating phase. Fig. 8 shows that the other outlier, at $t = N = 40$, is due to a peak in the probability mass function (pmf) of d_1, caused

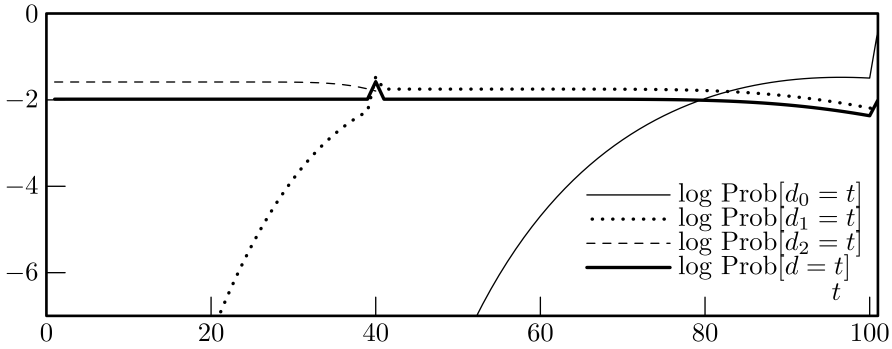

Fig. 7. Probability mass functions of the customer delays d_i $(i \in \{0, 1, 2\})$ and d on a log scale for $N = 40$, $T = 100$ and $\lambda = 0.4$.

by cycles in which the threshold N is reached. Of course, in every cycle where N is reached, the last customer will always have a delay N. Note however that all customers that arrive during consecutive slots share the same delay, so any customer that arrives in a series of consecutive slots with arrivals, that contains the phase's final slot, will have $d_1 = N$, thus creating the peak.

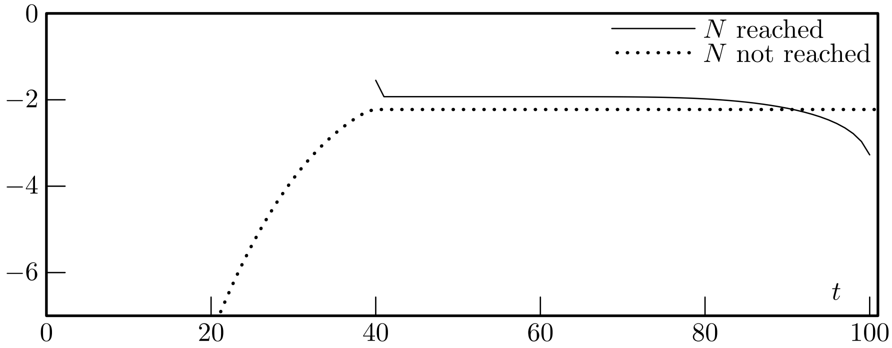

Fig. 8. Probability mass functions of the customer delay d_1 (split up) on a log scale for $N = 40$, $T = 100$ and $\lambda = 0.4$.

In Fig. 9, the effect of the arrival rate λ on the mean customer delay is presented, as well as the effect of λ on the weighted delays $p_iE[d_i]$ $(i \in \{0, 1, 2\})$. The graph shows that for extremely low arrival rates, the mean customer delay is dominated by the mean phase 0 customer delay $E[d_0]$. In accordance to p_0 the importance of d_0 rapidly decreases when the arrival rate increases, as a result of which the mean phase 1 delay becomes more important. Once the arrival rate is beyond $\frac{N}{T} = 0.4$, the N-threshold becomes ever more conclusive.

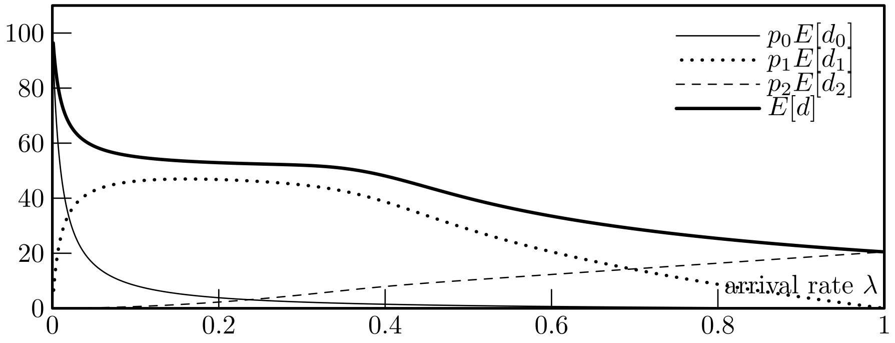

Fig. 9. Weighted mean customer delays $p_iE[d_i]$ $(i \in \{0, 1, 2\})$ and the overall mean customer delay $E[d]$ versus the arrival rate λ for $N = 40$ and $T = 100$.

8 Comparison with N-Policy

With respect to the basic N-policy, the NT-policy's main objective is to eliminate the possibility of starvation due to a low arrival rate of the customers. Indeed, due to the time threshold T no customer delay can ever exceed $T+1$, whereas under the N-policy there is no upper bound for the customer delay.

In Fig. 10, the mean customer delay for both the N-policy and the NT-policy is presented as a function of the system load $\rho = \lambda$. This shows clearly how the NT-policy has much better performance than the N-policy in case of a low rate arrival stream. For $\lambda > \frac{N}{T}$, there is only little benefit of the threshold T and both policies have an identical performance.

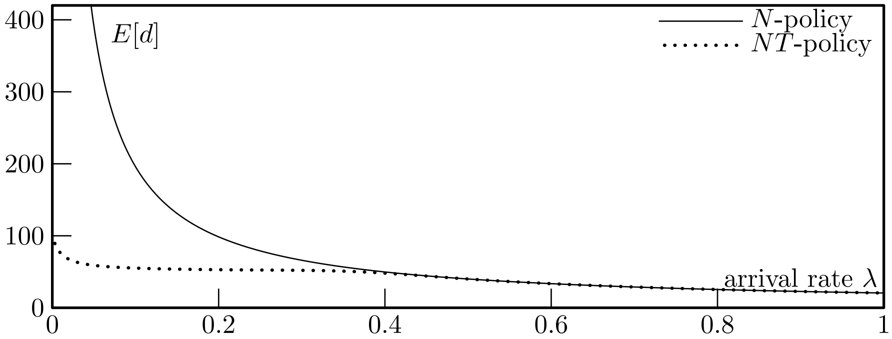

Fig. 10. Mean customer delay $E[d]$ for both N-policy and NT-policy versus the arrival rate λ for $N = 40$ and $T = 100$.

We should however note that the comparison presented here is somewhat unfair, since we assumed identical values for N in both policies. Parameter optimization for a specific cost model would, especially in the low rate traffic case, result in different values for N.

9 Conclusion

In this paper, we have studied the NT-policy in a discrete-time queueing system with independent Bernoulli arrivals and a deterministic server. We have obtained the distribution of the sojourn times of the three system phases. We also derived the customer delay distribution, conditioned on the phase during which the customer arrives. With some numerical examples, we illustrated the features and characteristics of the NT-policy. Finally, we compared the delay performance of the NT-policy with that of its more basic variant, the N-policy.

References

[1] M. Yadin and P. Naor, Queueing systems with a removable service station, *Operational Research Quarterly*, vol. 14, no. 4, pp. 393–405 (1963)

[2] S.S. Lee, H.W. Lee and K.C. Chae, Batch arrival queue with N-policy and single vacation, *Computers & Operations Research*, vol. 22, no. 2, pp. 173-189 (1995)

[3] K.-H. Wang, T.-Y. Wang and W.L. Pearn, Optimal control of the N-policy $M/G/1$ queueing system with server breakdowns and general startup times, *Applied Mathematical Modelling*, vol. 31, no. 10, pp. 2199–2212 (2007)

[4] J.-C. Ke, H.-I Huang and Y.-K. Chu, Batch arrival queue with N-policy and at most J vacations, *Applied Mathematical Modelling*, vol. 34, no. 2, pp. 451–466 (2010)

[5] W. Böhm and S.G. Mohanty, On discrete-time Markovian N-Policy queues involving batches, *Sankhya: The Indian Journal of Statistics, Series A*, vol. 56, no. 1, pp. 144–163 (1994)

[6] A.G. Hernández-Díaz and P. Moreno, Analysis and optimal control of a discrete-time queueing system under the (m, N)-policy, *Valuetools '06: Proceedings of the 1st international conference on Performance evaluation methodolgies and tools*, (Pisa, Italy, 2006)

[7] P. Moreno, A discrete-time single-server queue with a modified N-policy, *International Journal of Systems Science*, vol. 38, no. 6, pp. 483–492 (2007)

[8] J.-C. Ke, Optimal NT policies for M/G/1 system with a startup and unreliable server, *Computers & Industrial Engineering*, vol. 50, no. 3, pp. 248–262 (2006)

[9] H.W. Lee and W.J. Seo, The performance of the $M/G/1$ queue under the dyadic Min(N, D)-policy and its cost optimization, *Performance Evaluation*, vol. 65, no. 10, pp. 742–758 (2008)

[10] A.S. Alfa, W. Li, Optimal (N,T)-policy for $M/G/1$ system with cost structures, *Performance Evaluation*, vol. 42, no. 4, pp. 265–277 (2000)

[11] O.J. Boxma, W.P. Groenendijk, Waiting times in discrete-time cyclic-service systems, *IEEE Transactions on Communications*, vol. 36, no. 2, pp. 164-170 (1988)